Complete genome sequence of *Leptotrichia buccalis* type strain (C-1013-b^T)

Natalia Ivanova[1], Sabine Gronow[2], Alla Lapidus[1], Alex Copeland[1], Tijana Glavina Del Rio[1], Matt Nolan[1], Susan Lucas[1], Feng Chen[1], Hope Tice[1], Jan-Fang Cheng[1], Liz Saunders[1,3], David Bruce[1,3], Lynne Goodwin[1,3], Thomas Brettin[1,3], John C. Detter [1,3], Cliff Han[1,3], Sam Pitluck[1], Natalia Mikhailova[1], Amrita Pati[1], Konstantinos Mavromatis[1], Amy Chen[4], Krishna Palaniappan[4], Miriam Land[1,5], Loren Hauser[1,5], Yun-Juan Chang[1,5], Cynthia C. Jefferies[1,5], Patrick Chain[1,6], Christine Rohde[2], Markus Göker[2], Jim Bristow[1], Jonathan A. Eisen[1,7], Victor Markowitz[4], Philip Hugenholtz[1], Nikos C. Kyrpides[1], and Hans-Peter Klenk[2]*

[1] DOE Joint Genome Institute, Walnut Creek, California, USA
[2] DSMZ - German Collection of Microorganisms and Cell Cultures GmbH, Braunschweig, Germany
[3] Los Alamos National Laboratory, Bioscience Division, Los Alamos, New Mexico USA
[4] Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, USA
[5] Oak Ridge National Laboratory, Oak Ridge, TN, USA
[6] Lawrence Livermore National Laboratory, Livermore, California 94550, USA
[7] University of California Davis Genome Center, Davis, California, USA

[*]Corresponding author: Hans-Peter Klenk

Keywords
Fusobacteria, '*Leptotrichiaceae*', Gram-negative fusiform rods, human oral microflora, dental plaque, non-motile, non-sporulating, anaerobic

Abstract
Leptotrichia buccalis (Robin 1853) Trevisan 1879 is the type species of the genus, and is of phylogenetic interest because of its isolated location in the sparsely populated and neither taxonomically nor genomically adequately accessed family '*Leptotrichiaceae*' within the phylum '*Fusobacteria*'. Species of *Leptotrichia* are large fusiform non-motile, non-sporulating rods, which often populate the human oral flora. *L. buccalis* is anaerobic to aerotolerant, and saccharolytic. Here we describe the features of this organism, together with the complete genome sequence and annotation. This is the first complete genome sequence of the order '*Fusobacteriales*' and no more than the second sequence from the phylum '*Fusobacteria*'. The 2,465,610 bp long single replicon genome with its 2306 protein-coding and 61 RNA genes is a part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project.

Introduction
Leptotrichia buccalis strain C-1013-b^T (DSM 1135 = ATCC 14201 = JCM 12969, and other strain collections) is the type strain of the species [1], which also represents the type species of the genus first adequately and validly described in 1879 by Trevisan to accommodate the oral filamentous bacteria and to separate them from the algae [2, 3]. For a while two entirely different organisms have been termed *L. buccalis* in the literature [3]. One of these was '*Leptothrix buccalis*', a name originally employed by Robin in 1853 for filamentous forms which he had seen in wet mounts of tooth scrapings [4]. Over a century of the history of classification and misclassification of *L. buccalis* was documented by Gilmore *et al.* 1961 [3]. *L. buccalis* was among the first bacteria to be described and drawn in the letters of Antoni van

Leeuwenhoek [5]. Next to *Fusobacterium nucleatum* [6], *L. buccalis* is only the second species from the phylum *Fusobacteria* for which a complete genome sequence is described.

Here we present a summary classification and a set of features for *L. buccalis* strain C-1013-b^T (Tab. 1), together with the description of the complete genomic sequencing and annotation.

Classification and features of organism

The primary habitat of *L. buccalis* and most other *Leptotrichia* species is the human oral cavity, especially dental plaque, or the female genitourinary tract and the intestinal tract [7, 8]. Although *L. buccalis* and *L. buccalis*-like bacteria have also occasionally been recovered from blood mostly in immunocompromised patients, they are not known as causative agents of systemic infections [7, 9] even if an endotoxin was documented for the *L. buccalis* [5, 9]. Almost all of the cultivated *Leptotrichia* isolates cluster in 16S rRNA sequence comparisons with one of the five other type strains of the genus *Leptotrichia* [7, Fig. 1]. Except for the uncultured clone GI5-008-C04 (FJ192568), which has been recovered from screening of a spacecraft assembly clean room during the Phoenix mission, all significantly related phylotypes were from the usual habitats as described above. No phylotypes from environmental screening or genomic surveys could be linked with more than 85% 16S rRNA sequence similarity to *L. buccalis* (status May 2009).

Fig. 1 shows the phylogenetic neighborhood of *L. buccalis* strain C-1013-b^T in a 16S rRNA based tree. The sequences of the five 16S rRNA gene copies in the genome of strain C-1013-b^T differ from each other by 5 to 20 nucleotides (up to 1.3%), and by 4 to 16 nucleotides plus 38 ambiguities (total up to 3.6%) from the previously published 16S rRNA sequence (X90831) generated from NCTC 10429.

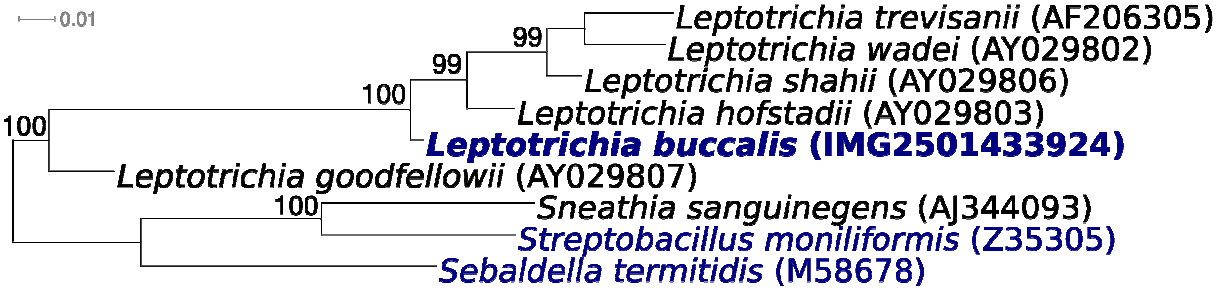
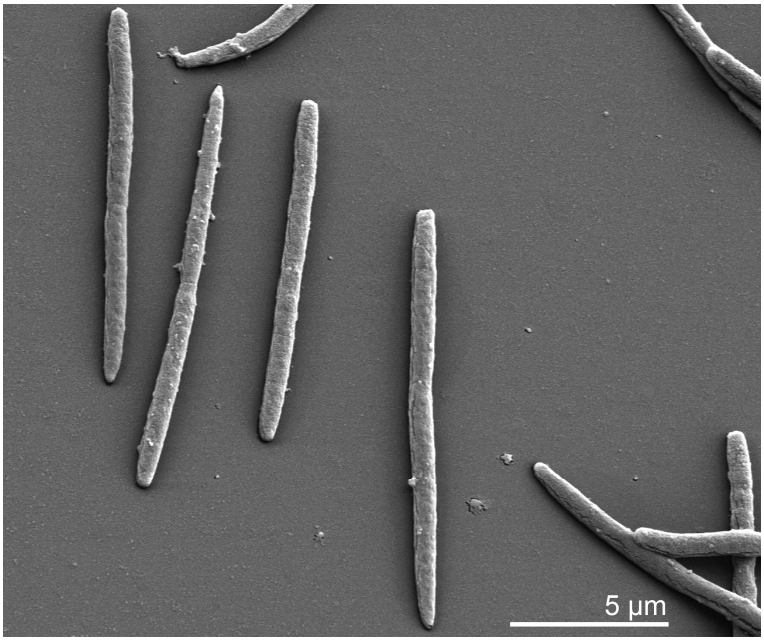


Figure 1. Phylogenetic tree highlighting the position of *L. buccalis* C-1013-b^T relative to all type strains of the genus *Leptotrichia* inferred from 1421 aligned characters [10, 11] of the 16S rRNA sequence under the maximum likelihood criterion [12], and rooted with all type strains of the family *Leptotrichiaceae*. The branches are scaled in terms of the expected number of substitutions per site. Numbers above branches are support values from 1000 bootstrap replicates if larger than 60%. Lineages with type strain genome sequencing projects registered in GOLD [13] are shown in blue, published genomes in bold.

Older cells of *L. buccalis* strain C-1013-b^T are gram negative, but younger cells that have been in culture for less than six hours show Gram-positiveness [5]. The organism forms long rods, commonly occurring in pairs, and is non-motile [5, Fig. 2]. Young colonies are colourless, smooth, shiny, raised and described as “medusa-head” colonies because of filamentous edges [5]. On first isolation, *L. buccalis* is anaerobic but becomes aerotolerant upon transfer and grows in the presence of air and CO_2 [5, 8]. *L. buccalis* is susceptible to many antibiotics but resistant to aminoglycosides [5]. The organism is highly saccharolytic and ferments a range of

different sugars [5, 8]. The main metabolic end product is lactic acid [8]. The G+C content was already in 1982 described as ‘unusually low’ (25 %) [5].

Figure 2. Scanning electron micrograph of *L. buccalis* stain C-1013-b^T (M. Rohde, HZI Braunschweig)



***Chemotaxonomy*.** The murein of strain C-1013-b^T contains *meso*-2,6-diaminopimelic acid (A_2pm), D- and L-alanine, and D-glutamic acid [14]. The strain possesses muramic acid and glucosamine as principal components of its peptidoglycan [14]; typeA1γ according to the classification of Schleifer and Kandler [15]. As in other *Leptotrichia* strains, the fatty acid pattern of *L. buccalis* is an almost equal mixture of saturated and unsaturated straight chain acids: $C_{16:0}$ (39%), $C_{14:0}$ (10%), $C_{18:1}$ (42%), with about 7% hydroxy acids ($C_{14:0}$) [7]. The type of menaquinones and polar lipids used by *L. buccalis*has not been described yet.

Table 1. Classification and general features of *L. buccalis* strain C-1013-b^T in accordance to the MIGS recommendations [16]

MIGS ID	Property	Term	Evidence code[a,b]
	Current classification	Domain *Bacteria*	
		Phylum *'Fusobacteria'*	
		Class *'Fusobacteria'*	
		Order *'Fusobacteriales'*	
		Suborder	
		Family *'Leptotrichiaceae'*	
		Genus *Leptotrichia*	TAS [2]
		Species *Leptotrichia buccalis*	TAS [2]
		Type strain C-1013-b	TAS [1]
	Gram stain	negative	TAS [5]
	Cell shape	long rods	TAS [5]
	Motility	nonmotile	TAS [5]
	Sporulation	nonsporulating	TAS [5]

	Temperature range	mesophile	NAS
	Optimum temperature	37°C	NAS
	Salinity	normal	NAS
MIGS-22	Oxygen requirement	anaerobic on isolation, becomes aerotolerant on further transfer	TAS [5]
	Carbon source	mono- and disaccharides	TAS [5]
	Energy source	carbohydrates	NAS
MIGS-6	Habitat	oral cavities	TAS [5]
MIGS-15	Biotic relationship	Free living	
MIGS-14	Pathogenicity	opportunistic pathogen	TAS [5]
	Biosafety level	1	TAS [17]
	Isolation	human oral flora	TAS [7]
MIGS-4	Geographic location	global	NAS
MIGS-5	Sample collection time	Mid of 19th century	TAS [3]
MIGS-4.1 MIGS-4.2	Latitude – Longitude	not reported	
MIGS-4.3	Depth	not reported	
MIGS-4.4	Altitude	not reported	

a) Evidence codes - IDA: Inferred from Direct Assay (first time in publication); TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from http://www.geneontology.org/GO.evidence.shtml of the Gene Ontology project [18]. If the evidence code is IDA, then the property should have been directly observed, for the purpose of this specific publication, for a live isolate by one of the authors, or an expert or reputable institution mentioned in the acknowledgements.

Genome sequencing information

Genome project history

This organism was selected for sequencing on the basis of its phylogenetic position, and is part of the ***G**enomic **E**ncyclopedia of **B**acteria and **A**rchaea* project. The genome project is deposited in the Genomes OnLine Database [13] and the complete genome sequence in GenBank NOT YET. Sequencing, finishing and annotation were performed by the DOE Joint Genome Institute (JGI). A summary of the project information is shown in Table 2.

Table 2. Genome sequencing project information

MIGS ID	Property	Term
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Three genomic libraries: two Sanger libraries - 8 kb pMCL200 and fosmid pcc1Fos and one 454 pyrosequence standard library
MIGS-29	Sequencing platforms	ABI3730, 454 GS FLX
MIGS-31.2	Sequencing coverage	9.7x Sanger; 42x pyrosequence
MIGS-30	Assemblers	Newbler version 1.1.02.15, phrap

MIGS-32	Gene calling method	Prodigal
	INSDC / Genbank ID	not yet available
	Genbank Date of Release	not yet available
	GOLD ID	Gi02240
	NCBI project ID	29445
	Database: IMG-GEBA	2501416906
	Project relevance	Tree of Life, GEBA

Growth conditions and DNA isolation
L. buccalis strain C-1013-b^T, DSM 1135, was grown anaerobically in DSMZ medium 104 (http://www.dsmz.de) at 37°C. DNA was isolated from 1-1.5 g of cell paste using Qiagen Genomic 500 DNA Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions, but with a modified protocol for cell lysis, using more lysozyme (1.6x) and an extended incubation time (1 hour).

Genome sequencing and assembly
The genome was sequenced using a combination of Sanger and 454 sequencing platforms. All general aspects of library construction and sequencing performed at the JGI can be found at http://www.jgi.doe.gov/. 454 pyrosequencing reads were assembled using the Newbler assembler version 1.1.02.15 (Roche). Large Newbler contigs were broken into 2,747 overlapping fragments of 1000bp and entered into assembly as pseudo-reads. The sequences were assigned quality scores based on Newbler consensus q-scores with modifications to account for overlap redundancy and to adjust inflated q-scores. A hybrid 454/Sanger assembly was made using the parallel phrap assembler (High Performance Software, LLC). Possible mis-assemblies were corrected with Dupfinisher or transposon bombing of bridging clones [19]. Gaps between contigs were closed by editing in Consed, custom primer walk or PCR amplification. 908 Sanger finishing reads were produced to close gaps, to resolve repetitive regions, and to raise the quality of the finished sequence. The error rate of the completed genome sequence is less than 1 in 100,000. Together all sequence types provided 51.7 x coverage of the genome. The final assembly contains 28,754 Sanger reads in addition to the 454 based pseudo reads.

Genome annotation
Genes were identified using Prodigal [20] as part of the Oak Ridge National Laboratory genome annotation pipeline, followed by a round of manual curation using JGI's GenePRIMP pipeline (http://geneprimp.jgi-psf.org) [21]. The predicted CDSs were translated and used to search the National Center for Biotechnology Information (NCBI) nonredundant database, UniProt, TIGRFam, Pfam, PRIAM, KEGG, COG, and InterPro databases. Additional gene prediction analysis and functional annotation was performed within the Integrated Microbial Genomes (IMG-ER) platform (http://img.jgi.doe.gov/er) [22].

Genome properties

The genome is 2,345,610 bp long and comprises one circular chromosome with a 29.7% GC content (Tab. 3). Of the 2367 genes predicted, 2306 were protein coding genes, and 61 RNAs; 86 pseudogenes were also identified. 65.4% of the genes were assigned with a putative function while the remaining are annotated as hypothetical proteins. The distribution of genes into GOGs functional categories is presented in Table 4.

Table 5. Genome Statistics

Attribute	Value	% of Total
Genome size (bp)	2,465,610	
DNA Coding region (bp)	2,139,206	86.76%
DNA G+C content (bp)	730,947	29.65%
Number of replicons	1	
Extrachromosomal elements	0	
Total genes	2367	100.00%
RNA genes	61	2.58%
rRNA operons	5	
Protein-coding genes	2306	97.42%
Pseudo genes	86	3.63%
Genes with function prediction	1547	65.36%
Genes in paralog clusters	402	16.98%
Genes assigned to COGs	1533	64.77%
Genes assigned Pfam domains	1577	66.62%
Genes with signal peptides	432	18.25%
Genes with transmembrane helices	530	22.39%
CRISPR repeats	4	

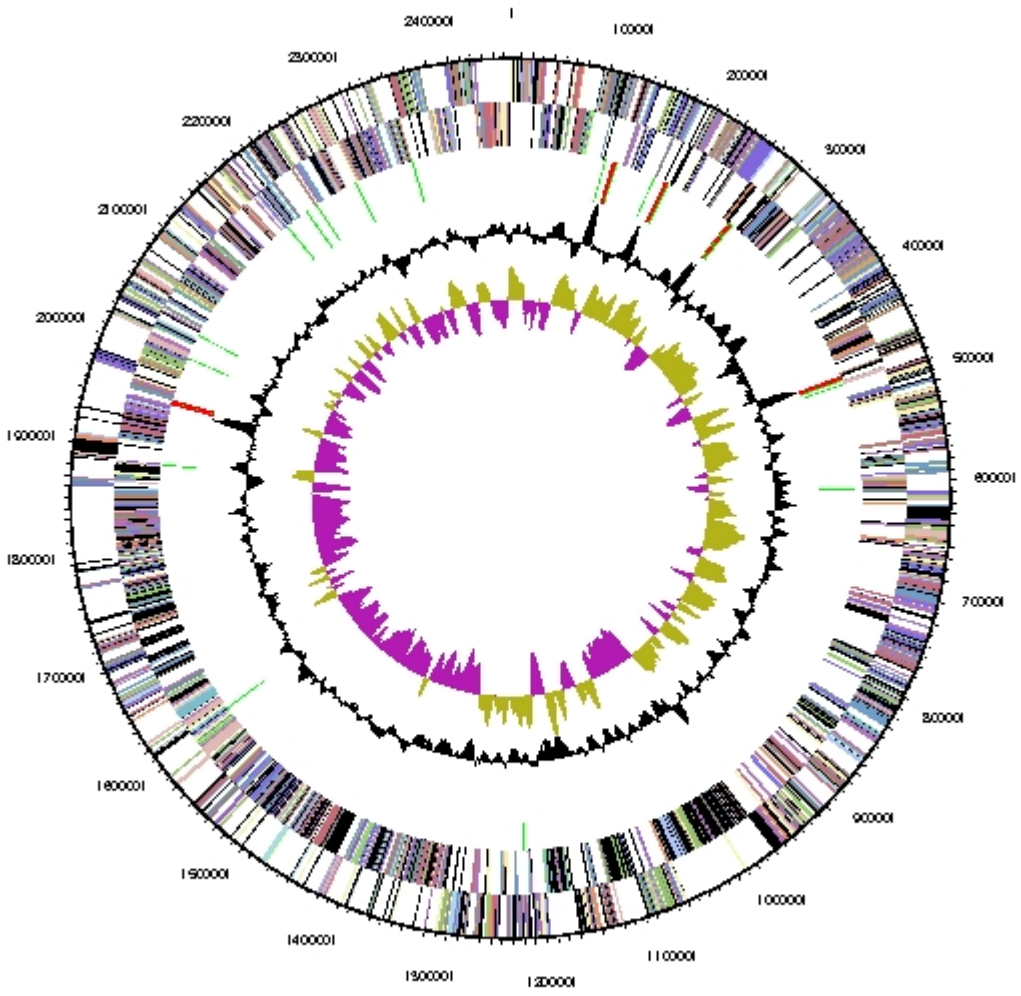


Figure 3. Graphical circular map of the genome. From outside to the center: Genes on forward strand (color by COG categories), Genes on reverse strand (color by COG categories), RNA genes (tRNAs green, sRNAs red, other RNAs black), GC content, GC skew.

Table 4. Number of genes associated with the 21 general COG functional categories

Code	COG counts and percentage of protein-coding genes		Description
	Genome		
	value	% of total	
J	144	6.2	Translation, ribosomal structure and biogenesis
A	0	0.0	RNA processing and modification
K	86	3.7	Transcription
L	124	5.4	Replication, recombination and repair
B	0	0.0	Chromatin structure and dynamics
D	24	1.0	Cell cycle control, mitosis and meiosis

Y	0	0.0	Nuclear structure
V	28	1.2	Defense mechanisms
T	47	2.0	Signal transduction mechanisms
M	112	4.9	Cell wall/membrane biogenesis
N	6	0.3	Cell motility
Z	0	0.0	Cytoskeleton
W	0	0.0	Extracellular structures
U	34	1.5	Intracellular trafficking and secretion
O	68	3.0	Posttranslational modification, protein turnover, chaperones
C	79	3.4	Energy production and conversion
G	110	4.8	Carbohydrate transport and metabolism
E	176	7.6	Amino acid transport and metabolism
F	54	2.3	Nucleotide transport and metabolism
H	80	3.5	Coenzyme transport and metabolism
I	44	1.9	Lipid transport and metabolism
P	84	3.6	Inorganic ion transport and metabolism
Q	11	0.5	Secondary metabolites biosynthesis, transport and catabolism
R	206	8.9	General function prediction only
S	141	6.1	Function unknown
-	773	28.1	Not in COGs

Acknowledgements

We would like to gratefully acknowledge the help of Sabine Welnitz for growing *L. buccalis* cultures and Susanne Schneider for DNA extraction and quality analysis (both at DSMZ). This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396, as well as German Research Foundation (DFG) INST 599/1-1.

References

1. Skerman VBD, McGowan V, Sneath PHA. Approved list of bacterial names. *Int J Syst Bacteriol* 1980, **30:**225-30.

2. Trevisan V. Prime linee d'introduzione allo studio dej Batterj italiani. *Rend Reale Inst Lombardo Sci, Ser II* 1879, **12:**133-51.

3. Gilmour MN, Howell A Jr, Bibby BG. The classification of organisms termed *Leptotrichia (Leptothrix) buccalis*. *Bacteriol Rev* 1961, **25:**131-41.

4. Robin C. Histoire naturelle des evégétaux parasites qui croissent sur l'homme et sur les animaux vivantis, JB Baillière, Paris 1853, pp 345-454.

5. Hofstad T. The Genus *Leptotrichia*. *In*: The Prokaryotes Vol. IV, 2nd ed. 1992, Springer-Verlag New York.

6. Kapatral V, Anderson I, Ivanova N, Reznik G, Los T, *et al*.. Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586. *J Bacteriol* 2002, **184:**2005-18.

7. Eribe ERK, Paster BJ, Caugant DA, Dewhirst FE, Stromberg VK *et al*.. Genetic diversity of *Leptotrichia* and description of *Leptotrichia goodfellowii* sp. nov., *Leptotrichia hofstadii* sp. nov., *Leptotrichia shahii* sp. nov. and *Leptotrichia wadei* sp. nov. *Int J Syst Evol Microbiol* 2004, **54**:583-92.

8. Hofstad T, Jantzen E. Fatty Acids of *Leptotrichia buccalis*: Taxonomic Implications. *J Gen Microbiol* 1982, **128**:151-3.

9. Bhally HS, Lema C, Borek A, Carroll KC. *Leptotrichia buccalis* bacteremia in two patients with acute myelogenous leukemia. *Infectious Diseases J of Pakistan* 2004, Apr-Jun:49-51.

10. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. *Bioinformatics* 2002, **18**:452-64.

11. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 2000, **17**:540-52.

12. Stamatakis A, Hoover P, Rougemont J. A rapid bootstrap algorithm for the RAxML web-servers. *Syst Biol* 2008, **57**:758-71.

13. Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC. The Genomes OnLine Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res* 2008, **36**:D475-9.

14. Fukuhara H, Umemoto T, Sagawa H, Kato K, Kotani S. Purification and quantitative chemical analysis of cell wall peptidoglycans of *Leptotrichia buccalis*. Infection and Immunity 1983, **39**: 132-6.

15. Schleifer KH, Kandler O. Peptidoglycan types of bacterial cell walls and their taxonomic implications. *Bacteriol Rev* 1972, **36:**407-77.

16. Field D, Garrity G, Gray T, Morrison N, Selengut J, *et al*. Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. *Nat Biotechnol* 2008, **26**:541-7.

17. Biological Agents: Technical rules for biological agents www.baua.de TRBA 466.

18. The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet* 2000, **25**:25-9.

19. Sims D, Brettin T, Detter JC, Han C, Lapidus A, *et al*. Complete genome of *Kytococcus sedentarius* type strain (strain 541T). *SIGS*, 2009 reviewed.

20. http://compbio.ornl.gov/prodigal/

21. Pati et al. GenePRIMP: A Gene Prediction Improvement Pipeline for microbial genomes. *in preparation* 2009.

22. Markowitz VM, Mavromatis K, Ivanova NN, Chen I-MA, Kyrpides NC. Expert Review of Functional Annotations for Microbial Genomes. *Bioinformatics* 2009, in press.